A NOVEL PROBABILISTIC BASED FEATURE SELECTION MODEL FOR CREDIT CARD ANOMALY DETECTION

[1]Y.A.SIVA PRASAD, [2]DR.G.RAMAKRISHNA

[1]Research Scholar, Department of Computer Science and Engineering, KL University, Andhra Pradesh

[2]Professor, Department of Computer Science and Engineering, KL University, Andhra Pradesh

E-mail: [1]sivaprasady@gmail.com, [2]ramakrishnag_cse@kluniversity.in

ABSTRACT

Due to the increase in online financial applications, the fraudulent operations through online transactions have increased rapidly. Also, the anomaly detection in credit card transactions has become equally important in many fields in which the data have high dimensional attributes. Finding noisy anomaly attributes using the conventional models are inefficient and infeasible, as the size and number of instances are large. In this paper, an optimized probabilistic based feature selection model was implemented on credit card fraud detection. An efficient ranked attributes are extracted using the hybrid feature selection algorithm. Experimental results show that proposed system efficiently detects the relevant attributes compared to traditional models in terms of time and dimensions are concerned

.

Keywords: *Feature selection algorithm, Fraud detection, Markov model, density distribution.*

1. INTRODUCTION

Class imbalance is a problem wherein the class distribution among instances is skewed towards negatively or positively. For example: In multi-project defect detection, one class has fewer instances compare to the other class instances distribution. High dimensionality defines the datasets which have a large number of independent features. In high dimensionality phase datasets, only a small percentage of attributes provides interesting information about the class, and the rest of them may be duplicated and inconsistent. In addition to this, detecting the most relevant features is often more challenging task for defect prediction process. Since the software quality estimation model is based on the software metrics of the SDLC phase, the selection of relevant metrics or dependency metrics becomes an integrated part of the model building process.

Traditional feature selection models can be classified into two groups, one is the feature subset selection and another one is feature ranking. In feature ranking models, each feature is assessed according to the computed measures and then an analyst selects relevant features for a given data set. A feature subset selection model extracts a subset of features from the large set of features using selection measures.

Anomaly detection is concerned with finding exceptional objects. Distributed Credit cards have become a essential financial assessment tool for its multi-functions of making transactions, depositing, consumer credit and back transfer along with withdrawal of cash etc. The main analysis or purpose of credit card usage is to classify the users into two groups, users with bad credit score and users with good credit score.

An anomaly is an observation that deviates so much from each observation. Many anomaly detection techniques have been proposed in different categories such as depth based, clustering based , density based and distance based act.

Most of the anomaly detection models are implemented to handle continuous dataset. Since they need preprocessing of attributed which transform numerical attributed to binary and categorical to number attributes. Also, most of the existing models are not directly preprocessed to categorical data. Different techniques based on similar filter techniques have been introduced in the literature to detect anomalies in the credit score transactions.

For detecting anomalies, different approaches have been used [2-5] . Credit card anomaly detection has been usually seen as a machine learning challenge where the objective is to correctly classify the credit card transactions as legal or illegal.

For data classification problem many accuracy measures have been used, most of which are related to the hit ratio, mutual information, gini index and lift are the most detection measures. Also, credit card usage varies widely across households due to a number of consumer behavior like band choice, product choice, dealer choice ,income group and purchase amount.

Finding anomalies in the credit card data is a challenging task due to high dimensional features, noise values and imbalance property.

The rest of this paper is organized as follows, Section II describes the literature study of different feature selection and classification models. Section III describe the chain based proposed feature selection models, Section IV describes the experimental results and in Section V, we conclude with the model.

2. RELATED WORK

Naïve bayes[1] is a very effective classification technique to predict the existence of anomaly based on the training samples. A naïve Bayes model considers anomaly prediction as a binary classifier i.e. it trains and predicts predictor by analyzing historical metric data. If the attribute types in the metric data are mixed type, then it is difficult to predict the anomalies due to missing values or uncertain data. Since, most of the features are categorical or binary attributes with normal distribution , it is difficult to predict the numerical attributes using prior and posterior probability values.

KNN[2] method to judge the anomaly rate in credit card status and events. They try to give the credit card anomaly rates using some statistic techniques. Since credit card dataset have multiple attribute types which are not possible to filter the attributes using KNN model. With the data mining techniques more mature and widely used, for analysis and mining the hidden information in the development repository become a hot research topic. The usual ways which use data mining techniques in this domain include Association Rules, Classification and Prediction, Clustering.

Deviations from the normal indicate anomalies that are then assumed to be an intrusion or attack. Different modeling approaches have included statistical methods, rule based systems, neural networks [3] and other soft-computing techniques [4].

Anomaly detection approaches can be classified into three main domain areas as shown in Figure 1. Statistical based anomaly detection systems and Knowledge based anomaly detection systems and

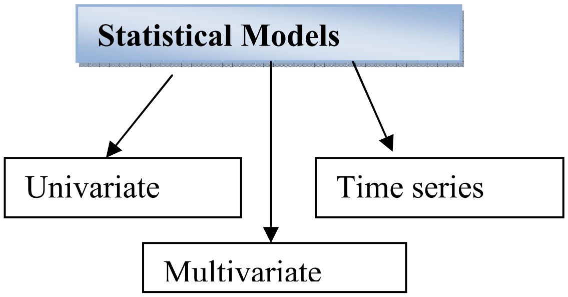


Fig 1: Statistical anomaly detection

Some of the major issues in these models should be pointed out: First, each model should be trained by an expert in such a way that the network packets generated during the attack are considered as normal type. Second, initializing the different values or metrics is a highly difficult task, especially in the case of false positive and true negative patterns are detected

Factors	Statistical Based IDS
Methodology	Observes statistical patternsof network traffic
Accuracy	Low
Speed	Fast
Limitations	Comprehensive training dataset required
Complexity	Low

.

2.1 Association Rule Mining Models

Association rule mining is one of the most widely used approaches in data mining technology and also used for network intrusion detection system. Association patterns define the relationship between the anomaly and normal features using the statistical support and confidence measures.There are two main phases to detect the network attacks using association rule mining techniques. Initially, it detects the candidate patterns using the credit card data and minimum support measure. Then, it constructs the frequent patterns using the minimum support and minimum confidence thresholds. These frequent patterns are evaluated to find the interesting relationships among the network features. Traditional models such as Apriori, FPgrowth, CPTree ,etc are used as network intrusion detection systems[5].

2.2 Sequential Pattern Analysis

Similar to association rule mining models , sequential models are designed for the purpose of mining the source of credit frauds and its associated root access with respect to time. There are two important features in the sequential mining process, such as the time gap between the transaction events and duration of the transaction patterns[6-7].

2.3 Clustering Models

The basic idea of clustering is to group the similar type of network patterns into meaningful subclasses so that the objects within the same cluster are most similar and the objects from different clusters are quite different from each other. Basically, clustering models are classified into four categories they are : hierarchical techniques, partitioning techniques ,grid based techniques and density based techniques[8-9].

Further, the more the anomaly data are, the higher the inter cluster entropy measure. Center based partitioning techniques such as k-means and k-medoids are the basic clustering methods due to their balancing and partitioning mechanism. K-medoids approach is more robust than traditional K-means based clustering algorithm.

Traditional anomaly detection model s discern frauds mainly depends on the database support and confidence measures which affects the overall true positive rate for detection process.

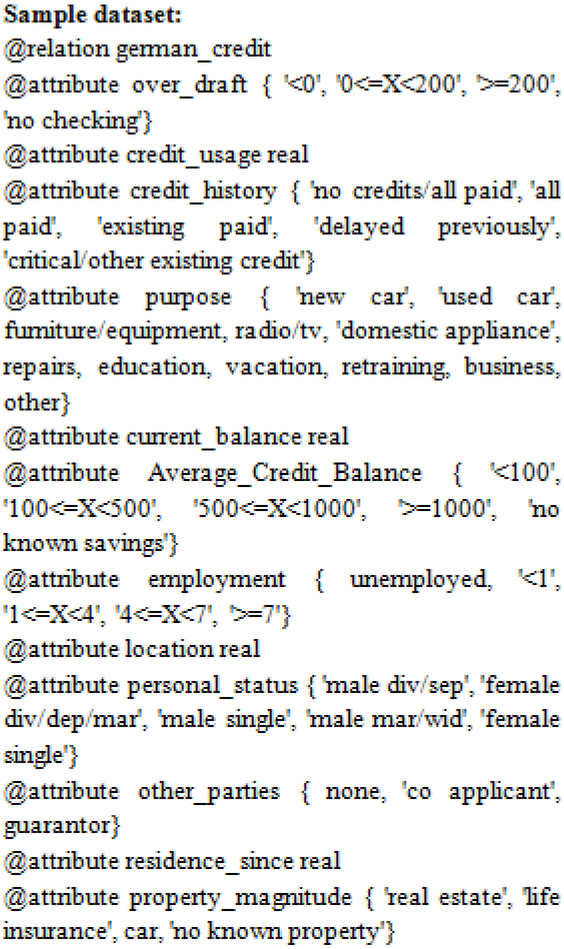
In this paper, credit card fraud detection model based on preprocessing, feature selection with similarity measure to predict the credit card anomalies in the given high dimensional features set are implemented.

3. PROPOSED MODEL

One of the major problems identified in the credit card data is the class imbalanced property of the distributed dataset. In order to handle efficiently the distributed credit card data such as online credit card anomaly detection applications requires large number of features rather than limited features. Dynamic analysis techniques, share the limitations of feature extraction inherently. Dynamic analysis cannot support complete analysis of target dataset features since it uses monitored partial behavior of the target features. The other limitation is that dynamic analysis techniques are difficult to be applied unless target features are balanced. Hence data should be preprocessed prior to the anomaly detection. The three phases included in the proposed framework are feature extraction, feature transformation and anomaly detection.

German Credit Dataset

The German Credit dataset has been given as training data with anomalies from the Machine learning UCI Repository.. This credit card dataset classifies users described by a set of features as good or bad as the credit risks



Sample dataset:

```
@relation german_credit
@attribute over_draft { '<0', '0<=X<200', '>=200', 'no checking'}
@attribute credit_usage real
@attribute credit_history { 'no credits/all paid', 'all paid', 'existing paid', 'delayed previously', 'critical/other existing credit'}
@attribute purpose { 'new car', 'used car', furniture/equipment, radio/tv, 'domestic appliance', repairs, education, vacation, retraining, business, other}
@attribute current_balance real
@attribute Average_Credit_Balance { '<100', '100<=X<500', '500<=X<1000', '>=1000', 'no known savings'}
@attribute employment { unemployed, '<1', '1<=X<4', '4<=X<7', '>=7'}
@attribute location real
@attribute personal_status { 'male div/sep', 'female div/dep/mar', 'male single', 'male mar/wid', 'female single'}
@attribute other_parties { none, 'co applicant', guarantor}
@attribute residence_since real
@attribute property_magnitude { 'real estate', 'life insurance', car, 'no known property'}
```

@attribute cc_age real

@attribute other_payment_plans { bank, stores, none}

```
@attribute housing { rent, own, 'for free'}
@attribute existing_credits real
@attribute job { 'unemp/unskilled non res',
'unskilled resident', skilled, 'high qualif/self
emp/mgmt'}
@attribute num_dependents real
@attribute own_telephone { none, yes}
@attribute foreign_worker { yes, no}
@attribute class { good, bad}
@data
'<0',6,'critical/other existing credit',radio/tv,1169,'no
known savings','>=7',4,'male single',none,4,'real
estate',67,none,own,2,skilled,1,yes,yes,good
'0<=X<200',48,'existing
paid',radio/tv,5951,'<100','1<=X<4',2,'female
div/dep/mar',none,2,'real
estate',22,none,own,1,skilled,1,none,yes,bad
'no checking',12,'critical/other existing
credit',education,2096,'<100','4<=X<7',2,'male
single',none,3,'real estate',49,none,own,1,'unskilled
resident',2,none,yes,good
'<0',42,'existing
paid',furniture/equipment,7882,'<100','4<=X<7',2,'
male single',guarantor,4,'life insurance',45,none,'for
free',1,skilled,2,none,yes,good
paid',furniture/equipment,7882,'<100','4<=X<7',2,'
male single',guarantor,4,'life insurance',45,none,'for
free',1,skilled,2,none,yes,good
'<0',42,'existing
paid',furniture/equipment,7882,'<100','4<=X<7',2,'
male single',guarantor,4,'life insurance',45,none,'for
free',1,skilled,2,none,yes,good
'<0',24,'delayed previously','new
car',4870,'<100','1<=X<4',3,'male single',none,4,'no
known property',53,none,'for
free',2,skilled,2,none,yes,bad
'no checking',36,'existing paid',education,9055,'no
known savings','1<=X<4',2,'male single',none,4,'no
known property',35,none,'for free',1,'unskilled
resident',2,yes,yes,good
'no checking',24,'existing
paid',furniture/equipment,2835,'500<=X<1000','>=
7',3,'male single',none,4,'life
insurance',53,none,own,1,skilled,1,none,yes,good
```

. Totally it contains 1500+ records in which there are 7 numerical features and 13 nominal features. Making it a total of 21 features together with class labels as Good and Bad.

Feature Generation

Generating set of features that reflect the underlying facts appropriately is not a trivial task. We extensively searched for features from the literature and selected features that are often used in the past researches. Further, we generated features using our own background knowledge. New features can be constructed by transforming or combining the original attributes. This approach is known as feature construction. It is often done by incorporating expert's background knowledge about the problem domain.

Feature Selection

Selecting the most suitable set of attributes that represent a problem, from large set of attributes is also a challenging task. Some attributes might be irrelevant, redundant, or containing useful information only when combined together. We must select the best possible features before feeding them into the algorithm since this influence the quality of the prediction model as well as the computer resources (such as calculation time, memory usage *etc.*).

In feature selection, the wrapper is the model evaluation based on different feature combinations. The evaluation result (e.g. the accuracy from a 10-fold cross validation) allows the identification of the best-performing model and thus, the best-performing feature combination. So, the best performing feature combination is the feature combination to select from all features.

There are three decisions to make to perform this kind of feature selection. First, what is the selection criterion to apply. Typically, the outcome of a classifier evaluation is the accuracy or the area under the ROC curve AUC [6]. These measures are the mostly used selection criteria following the rule: the higher, the better. Second, which algorithm to use. Although, the wrapper approach is concerned to be a black box approach to score the feature sub-sets, the algorithm choice has some influence on the results of the final model.

The algorithm used by the wrapper has less discriminative power than the subsequent learner and thus, unintentionally, omits valuable information. Third, we have to determine the appropriate search strategy. Ideally, wrapper methods would make use of all possible feature combinations to determine the feature contributions (exhaustive, complete search).

In feature selection, there are two fundamental search procedures, the forward and backward selection. Forward selection starts from scratch and adds new variables one-by-one while evaluating the optimal search path. The backward selection does the opposite: the search starts from a model based on all variables and eliminates one-by one. The results of both approaches can differ due to non-independent variables and different stopping points when a certain quality threshold value is reached. In other wrapper application fields also other search techniques such as evolutionary search and simulated annealing are used.

Classification is one of the essential techniques for software defect detection. Detection of defects can be performed using software features or attributes. Existing models for software defect pattern analysis, such as logistic regression, feed forward neural networks and fuzzy based discriminant analysis. The decision tree model can be interpreted by analyzing the tree structured patterns.

Main Objectives of this model:

- Remove noise in the credit card dataset using proposed data preprocessing model.
- Multivariate anomaly patterns with complex relationship.
- Handle mixed data-type and uncertain decisions.

3.1.Data Preprocessing algorithm

```
Database D,
For each data record in D
Do
For each feature F in the record
Do
        If(F!=NULL)
        Then
                Continue;
        Else
                F_type=check_type(F);
        If(F_type==numerical)
        Then
Miss_Value= $\frac{Max(F)*\sigma_F^2 - Min(F)*\mu^2_F}{N(N-1)*[\mathrm{Max(F)} - \mathrm{Min(F)}]}$
Value(F)=Miss_Value;
        End if
        If(type==Categorical)
        Then
        Freq[]=frequency(F);// each category of class attribute.
Probability of each instance value per class.
        Prob[]= $\sum_{i=1}^{m} \Pr ob(x_j / C_i)$ ;
                i=1,2,3…m classes
                j=1,2…n instances
rank=Max{freq[]}/Max{Prob[]};
Fill the value with the max ranked class value.
        End if
Done
Done
```

In this algorithm, missing values or inconsistent values are replaced with the computed value. If the attribute is numerical then all the missing values are replaced with the computed Max-Min value. If the attribute is categorical , then all the missing values are replaced with probabilistic ranked value.

3.2. Fraud detection attribute selection algorithm

```
Input : Filtered data FDB
Output: Ranked feature attributes
Procedure:
For each filtered feature ff in FDB
Do
Compute entropy E(ff);
Compute mutual information between attributes.
MI(ff)=Max{MI{ff,F-ff}};
Partition the feature ff into m classes as
Find the similarity between instances of two distinct partitions as
```

$$\mathrm{Sim}(p_i, p_j) = \frac{2 * \sum_{i,j} | x_i - x_j |^2}{N_i(N_j - 1)}$$

Where N_i is the number of instances in ith partition and N_j is the number of instances in jth partition.

Rank of the attribute is defined as

```
$R(ff)=E(ff)+M.I(ff)+Max\{ Sim(p_i, p_j) \}$

Done
Input k as user defined threshold
For each r in R(ff) do
If(r>k)
Then
Select as fraud feature attribute.
Else
Continue;
Done
```

In this algorithm, rank based fraud detection attributes are selected using a novel approach. In this model, entropy and mutual information measures are computed to each attribute with the remaining attributes. Also, similarity measure is computed to all the data partitions for intra cluster variations. The rank of an attribute is computed using the entropy, mutual information and similarity measure. Feature attributes are selected using the user defined threshold.

4. EXPERIMENTAL RESULT

All experiments were performed with the configurations Intel(R) Core(TM)2 CPU 2.13GHz, 4 GB RAM, and the operating system platform is Microsoft Windows XP Professional (SP2). There are two types of classes in credit card fraud detection process: one is fraudulent transactions and the other is non-fraudulent transactions.

In our experimental results can be divided into three phases:

1. Input a dataset of credit card transaction dataset with n features.
2. Computing the preprocessing and feature selection models for anomaly detection.
3. Finding the anomalies using the given similarity measure and the threshold.

Feature selection based fraud patterns:

employment != >=7 -> property_magnitude != no known property

foreign_worker != no AND personal_status != female single -> own_telephone != yes

other_payment_plans != none -> personal_status != female single

job != high qualif/self emp/mgmt -> purpose != other

foreign_worker != no -> personal_status != male mar/wid

property_magnitude != no known property -> existing_credits <= 4.0

personal_status != male mar/wid -> foreign_worker != no

current_balance <= 18424.0 -> class != bad

own_telephone != yes -> existing_credits <= 4.0

class != bad -> employment != >=7

other_parties != guarantor -> own_telephone != yes

existing_credits >= 1.0 AND foreign_worker != no -> other_payment_plans != none

property_magnitude != no known property AND own_telephone != yes -> purpose != other

existing_credits <= 4.0 -> other_parties != guarantor

current_balance <= 18424.0 -> purpose != other

housing != for free -> personal_status != female single

own_telephone != yes -> existing_credits >= 1.0

current_balance <= 18424.0 -> personal_status != female single

purpose != other -> job != high qualif/self emp/mgmt

other_parties != guarantor -> credit_history != critical/other existing credit

own_telephone != yes AND foreign_worker != no -> other_payment_plans != none

current_balance <= 18424.0 AND personal_status != female single -> foreign_worker != no

own_telephone != yes AND other_payment_plans != none -> personal_status != female single

other_payment_plans != none -> purpose != other

existing_credits <= 4.0 AND personal_status != female single -> foreign_worker != no

credit_history != critical/other existing credit -> employment != >=7

existing_credits >= 1.0 -> property_magnitude != no known property class != bad AND personal_status != female single -> foreign_worker != no

Number of Iterations :7

F-Measure: 0.82755

Recall : 0.939

TP rate : 0.966

FP rate : 0.0275

Classification Accuracy 0.9568

Table 1: Number of instances , attribute with computed ranked attributes

Number of instances	Number of features	Ranked Attributes
500	10	6
700	14	8
900	16	11
1200	18	15
1500	21	14

Table 1 describes the number of instances and its ranked attributes. These ranked attributes are used to find the anomalies and its patterns.

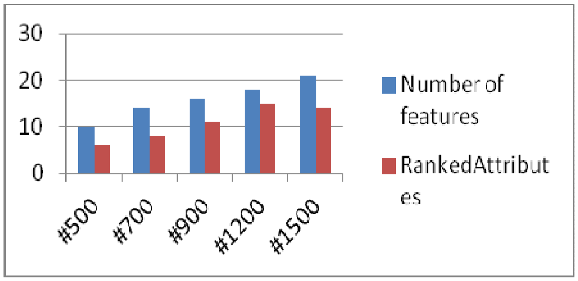


Fig 2: Comparison of the number of features and selected ranked attributes

Figure 2 describes the number of instances and its ranked attributes. These ranked attributes are used to find the anomalies and its patterns.

TABLE 2: Accuracy Comparison Of Proposed And Existing Algorithms

#instances	Naïve Bayes[1]	Neural Network[2]	Multi-objective [3]	Feature selection Based Multiobjective
500	0.92	0.87	0.893	0.945
700	0.89	0.85	0.905	0.935
900	0.9	0.87	0.9154	0.946
1200	0.913	0.834	0.912	0.953
1500	0.915	0.89	0.92	0.964

Table 3, describes the comparison of the existing and proposed models in terms of true positive and precision are concerned. From the table it is observed that proposed model has high computational rate compared to traditional models.

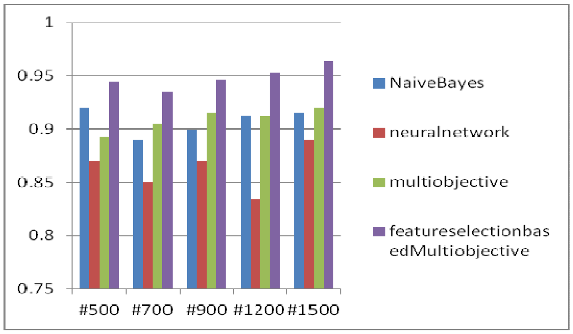


Fig 3: Accuracy Comparison Of Proposed And Existing Algorithms

Figure 3, describes the comparison of the existing and proposed models in terms of true positive and precision are concerned. From the Figure , it is observed that proposed model has high computational rate compared to traditional models.

5. CONCLUSION

In this paper, a novel probabilistic based feature selection model was implemented on credit card fraud detection. Finding noisy anomaly attributes using the conventional models are inefficient and infeasible, as the size and number of instances are large. In this paper, we have analyzed the anomalies in the credit card data through feature selection method. Experimental results show that proposed system efficiently detects the relevant attributes compared to traditional models in terms of time and dimensions are concerned. From this work, the true positive rate of the patterns are also analyzed to detect the fraudulent transactions in the test instances. In future, this work can be extended to detect the anomaly patterns in real-time web transactions.

REFRENCES:

[1] AbhinavSrivastava, AmlanKundu, ShamikSural and ArunK. Majumdar(2008) 'Credit Card Fraud Detection UsingHidden Markov Model' IEEE Transactions onDependable and Secure Computing vol. 5 No. 1.

[2] Bhattacharya.S, Jha.S, Tharakunnel. k, and Westland J.C, “Data mining for credit card fraud: A comparative study.” Decision Support systems, Vol.50, no. 7, 2011, pp.602-613.

[3] Clifton Phua, Vincent Lee, Kate Smith, Ross Gayler, ‘A Comprehensive Survey of Data Mining-based Fraud DetectionResearch’, http://arxiv.org/ftp/arxiv/papers/1009/1009.6119 .pdf.

[4] Dipti Thakur and Shalini Bhatia (2009) ‘Distributed Data Mining Approach to Credit Card Fraud Detection’ Proceedings of SPIT-IEEE .

[5] Aditya A. Davale; Shailendra W. Shende,"Implementation of coherent rule mining algorithm for association rule mining",Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015 International Conference on 2015.

[6]Quanzhong Liu; Yang Zhang; Zhengguo Hu,"Extracting Positive and Negative Association Classification Rules from RBF Kernel",IEEE,2010"

[7]Zhixin Hao; Xuan Wang; Lin Yao; Yaoyun Zhang,"Improved classification based on predictive association rules",IEEE,2009.

[8]Nitendra Kumar Vishwakarma; Jatin Agarwal; Sankalp Agarwal; Shantanu Sharma," Comparative analysis of different techniques in classification based on association rules",IEEE,2013

[9]Yanbo J. Wang; Qin Xin; Frans Coenen ,"A Novel Rule Weighting Approach in Classification Association Rule Mining",IEEE,2007